Ksenia Gnevsheva*

The expectation mismatch effect in accentedness perception of Asian and Caucasian non-native speakers of English

https://doi.org/10.1515/ling-2018-0006

Abstract: Previous research on speech perception has found an effect of ethnicity, such that the same audio clip may be rated more accented when presented with an Asian face (Rubin, Donald L. 1992. Nonlanguage factors affecting undergraduates' judgments of nonnative English-speaking teaching assistants. *Research in Higher Education* 33(4). 511–531. doi: 10.1007/bf00973770). However, most previous work has concentrated on Asian non-native English speakers, and Caucasian speakers remain under-explored. In this study, listeners carried out an accentedness rating task using stimuli from first language Korean, German, and English speakers in 3 conditions: *audio* only, *video* only, and *audiovisual*. Korean speakers received similar accentedness ratings regardless of condition, but German speakers were rated significantly less accented in the video condition and more accented in the audiovisual condition than the audio one. This result is explained as an expectation mismatch effect, whereby, when the listeners saw a Caucasian speaker they did not expect to hear a foreign accent, but if they actually heard one it was made more salient by their expectation to the contrary.

Keywords: Speech perception, sociophonetics, foreign accentedness, ethnicity, variation

1 Introduction

There is a growing body of research on sociophonetic variation in speech perception (see Campbell-Kibler 2010; Drager 2010; for a review). Previous studies have shown a strong link between social and linguistic information such that, on the one hand, the perceived linguistic information influences the listeners' assumptions about the speaker's social information (e.g., Campbell-Kibler 2007) and, on the other hand, assumed social information may affect how linguistic information is perceived (e.g., Hay et al. 2006a; Hay et al. 2006b; Niedzielski 1999; see discussion below).

***Corresponding author: Ksenia Gnevsheva,** School of Literature, Languages and Linguistics, Australian National University, Canberra ACT 2600, Australia, E-mail: ksenia.gnevsheva@anu.edu.au

The existing relationship between social and linguistic information may be explained by usage-based models of speech perception. Exemplar theory (e.g., Johnson 1997; Johnson 2006; Pierrehumbert 2003) suggests that our brain stores clouds of representations, *exemplars*, for a given category; and that these clouds are updated constantly through the perception-production loop. Sociolinguistic information, such as different speaker characteristics (sex, age, origin, etc.) and contextual information (speech styles, etc.), is associated and stored together with these exemplars. The sociolinguistic information is activated when associated exemplars are activated, and it can activate associated exemplars when it is accessed (Hay et al. 2006a: 370).

In non-native speakers, previous research has also shown that assumed social information may affect perceived linguistic information, such as degree of accent (e.g., see a discussion of Rubin 1992 below). An accent is a "cumulative auditory effect of those features of pronunciation which identify where a person is from, regionally or socially" (Vishnevskaya 2008: 235). When listeners hear substantial differences in a speaker's production, they can identify the person as coming from a different background. A specific example of that would be the accent of a non-native speaker (NNS), whose first language is different from that of the listeners, which I will call here a foreign accent. In perception studies, accentedness is conceptualized as a subjective measure of how strong listeners rate a speaker's accent.

A usage-based account is potentially insightful when considering not just the first language (L1) perception studies discussed above, but also studies of second language (L2) variation, including foreign-accentedness rating tasks. Usage-based models would predict that in a foreign accentedness rating task the items that activated the representations most similar to the ones associated with the listener or other native speakers would be judged as less foreign-accented whereas the items different from them or similar to the representations that have previously been identified as foreign-accented would be judged as more foreign-accented.

Accentedness perception has been shown to be influenced by a number of stimulus-independent factors (Kraut and Wulff 2013; Levi et al. 2007). A speaker's perceived ethnicity is one of such factors influencing his/her perceived accentedness. Anecdotes of native English speakers of a non-white background being perceived to have an accent are abundant: in Lippi-Green (1997) a monolingual English-speaking woman of Asian Indian decent is asked by a shopkeeper to speak more slowly because of her "accent". Several studies have explored the effect of ethnicity on reverse linguistic stereotyping, which is the idea that perceived social characteristics may influence perceived linguistic characteristics (Kang and Rubin 2009). Rubin (1992) and McGowan (2015) have

shown that the assumed ethnicity of the speaker influenced their perceived accentedness and intelligibility ratings, respectively. Rubin (1992) showed that the same native speaker (NS) of American English was rated as more accented when the listeners were presented with a picture of an Asian woman compared to when they were presented with a picture of a Caucasian woman, and this was attributed to listeners' negative bias. That is, listeners' negative bias towards Asian faces was said to influence their accentedness rating even when presented with audio stimuli from a native speaker with a Standard American English accent. McGowan (2015) finds a different effect, such that when presented with Asian-accented speech stimuli, listeners' intelligibility ratings are higher when they are shown a picture of an Asian face rather than a Caucasian face. This is not seen to be an effect of negative bias, then, but instead an effect of matching the types of stimuli: an audio clip from a native speaker of Standard American English mismatched with a picture of an Asian face increases accentedness ratings, but an audio clip from an Asian-accented speaker matched with a picture of an Asian face increases intelligibility scores.

These studies provide us with important information about how a speaker's assumed ethnicity can impact upon listeners' reaction to speech stimuli, but they both focus on a minority ethnicity (i.e. Asian speakers in the USA). Much less is known about the perceptual effects of presenting listeners with foreign-accented speech along with pictures of Caucasian faces. Specifically, little is known about the perceptual effects of reverse linguistic stereotyping when listeners are presented with visual stimuli of a Caucasian speaker but audio stimuli of a non-standard variety, such as foreign-accented speech. The effects discussed by Rubin (1992) and McGowan (2015) offer contradictory predictions for the accentedness rating of foreign-accented speech presented with a Caucasian speaker:

- If reverse linguistic stereotyping works in the same way as reported in Rubin (1992), we might expect a picture of a Caucasian face to reduce the accentedness rating of a foreign-accented audio clip.
- If the mismatch effect found by McGowan (2015) for intelligibility also underlies accentedness ratings, we would expect a picture of a Caucasian face to increase the ratings of a foreign-accented clip.

This study examines this issue by exploring the effect of visual input and speaker ethnicity on accentedness rating in foreign-accented speech produced by Asian and Caucasian second language speakers of English. I begin in the following section by discussing sociophonetic work which has considered how linguistic and social information is thought to be related, and the effect that one type of information can have on perceptions of the other. Then I specifically

address work in accentedness perception, elaborating on the studies introduced above. After presenting the method used in this study, I discuss the results in relation to reverse linguistic stereotyping and listener expectation.

2 Background

2.1 The inter-relationship between linguistic and social information

Many studies have shown that the way a person speaks affects listeners' perception of the speaker in terms of a range of social categories, in a form of linguistic stereotyping. For example, Campbell-Kibler (2007) found that two speech samples that differed only in the speaker's production of the (ING) variable were associated with different social categories: *–in* was associated more with lack of education, masculinity, and the country, while *–ing* was perceived to be more educated, gay, and urban.

The reverse has also been attested: perceived phonetic information has been found to be influenced by (assumed) social information, such as geographical region (Hay et al. 2006a; Niedzielski 1999), and the socio-economic status and age of the speaker (Hay et al. 2006b). Niedzielski (1999) found that the information the listeners were given about the origin of the speaker influenced their responses in a perception task. If listeners were told that the speaker was from Michigan, they did not choose the tokens that actually matched the speaker's vowel production but those that matched most closely with the listeners' expectations of Michigan speech. Hay et al. (2006a) found a similar effect of mentioning a geographical region with a population of listeners from New Zealand. Two groups of listeners were asked to choose a synthesized vowel which was most similar to that of the speaker's actual production, and mark it on an answer-sheet which had either "Australian" or "New Zealander" written at the top. All listeners heard the same speaker of New Zealand English (NZE), but chose synthesized vowels which were more similar to Australian English if their answer sheet had "Australian" at the top. Hay and Drager (2010) found the same effect when no region was explicitly mentioned but the listeners were shown stuffed toy kangaroos or koalas, associated with Australia, or stuffed toy kiwis, associated with New Zealand. They argued that once a region is primed, it can have a perceptual effect in the listening task.

Similar effects have been found with other social factors, such as socioeconomic class and age. Hay et al. (2006b) manipulated the perceived social class

and age of the speakers in a vowel identification task and presented listeners with audio input containing /iə/ and /eə/, which are merged for some speakers of NZE. They found a connection between the assumed social characteristics of the speaker and listener accuracy at identifying the produced vowel. Interestingly, Strand (2000) found a similar effect for sex where listeners recognized words more slowly when the pitch in the recording was atypical of speaker sex.

Hay and colleagues explain their findings by usage-based models of speech perception. Hay et al. (2006b: 479) suggest a relationship between identification accuracy and the difference between expected and actual production. When both the linguistic and sociolinguistic information is available and is congruent, this may facilitate access through more focused activation of representations, resulting in fewest identification errors. An expectation mismatch or incongruence between the actual production and the expected production, which comes to be expected because of what the listeners are told about the speakers, may lead to higher error rates (and/or slower reaction) as the mismatch between the perceived phonetic and social information would result in a more spread-out activation of representations and may inhibit access.

One can hypothesize that activation of experience-based representations with conflicting phonetic or social information at the same time may influence a listener's ratings. In the next section, I review existing work on foreign accentedness perception and its relationship to social information, namely ethnicity, in order to consider this hypothesis in more detail.

2.2 Foreign accentedness perception and ethnicity

A number of studies have explored the way assumed ethnicity of the speaker influences his/her perceived accentedness and intelligibility. As introduced above, reverse linguistic stereotyping has been explored by studies using visual stimuli, such as pictures of people of different ethnicities, to represent the speaker in accentedness rating tasks. In Rubin (1992) the same audio-recording of a native speaker of Standard American English was presented to students in a class with two different pictures supposedly representing the speaker: a Caucasian and an Asian woman. The students who were presented with a picture of an Asian woman rated the recording as more accented because they expected it to be accented, Rubin (1992) argues. Moreover, comprehension scores of listeners presented with an Asian picture were lower than of those presented with a Caucasian picture. This is a persuasive example of what the effect of visual stimuli might be on the perception of native-like linguistic input. However, it remains unclear what the accentedness ratings would have been if

the listeners had been presented with Asian and Caucasian faces matched with accented speech rather than Standard American English.

In an experiment involving foreign- and standard- accented speech, Yi et al. (2013) collected native English speaker (NES) listeners' intelligibility and perceived accentedness ratings of native speakers of American English and non-native English speakers (NNESs) of Korean L1 in audio only and audiovisual conditions. In the intelligibility experiment, word recognition in noise was better for NSs than NNESs and better in the audiovisual than audio condition; there was also a significant interaction such that audiovisual benefit was larger for NSs than for NNESs. In the accentedness rating experiment, six NS listeners were presented randomly and rated on a 9-point Likert scale 40 target sentences, each spoken by the 4 speakers (2 NS + 2 NNS) in the 2 conditions (audio and audiovisual), resulting in a total of 320 presentations. In line with predictions of the negative bias hypothesis, the authors argue, the Korean speakers were rated significantly more accented in the audiovisual condition than in the audio only condition, exhibiting an effect of ethnicity. However, it could be that the experiment design may have had an impact on the obtained results because, besides the small number of listener and speaker participants, the use of audio/audiovisual condition as a within-listener factor may have prompted the participants to notice the importance of the visual cue. This interpretation is supported by Yi et al. (2014)'s finding of a null condition effect in a clarity rating task with the same stimuli from the Yi et al. (2013) study. Following a similar method as Yi et al. (2013), Han-Gyol et al. (2014) find no significant effect of condition or an interaction between condition and speaker group suggesting that listeners found Korean speakers equally comprehensible in both audio only and audiovisual conditions.

Findings consistent with the reverse linguistic stereotyping account are explained differently by Babel and Russell (2015). In line with previous work, they demonstrate that an Asian native speaker of English is perceived to be more accented and less intelligible, but they argue that this is a reflection of listeners' prior experiences which result in an expectation to hear foreign-accented speech from Asian speakers. They find no relationship between the participants' bias scores and perception ratings, making stereotyping an unlikely explanation for such behavior.

McGowan (2015) explores intelligibility and perceived accentedness in Standard American English and foreign-accented speech. In the intellibility experiment, listeners were presented with foreign-accented speech together with an Asian or a Caucasian photograph or a silhouette. The listeners, who had a task of transcribing Chinese-accented speech presented in multi-talker babble noise, were found to be significantly more accurate when presented with

an Asian photograph than a Caucasian face, possibly due to a "mismatch-induced inhibition" in the latter. According to usage-based models of speech perception, Chinese-accented speech and an Asian picture together would activate a more focused set of experience-based representations enhancing intelligibility while a misalignment between the audio and visual input would result in a mismatch of expectations and spread activation more thinly inhibiting intelligibility. Although this was not an accentedness rating experiment, McGowan (2015) argues against the negative bias hypothesis as he found that socioindexical cues enhanced perception.

Reverse linguistic stereotyping and expectation mismatch predict conflicting outcomes in regard to foreign-accented Asian and Caucasian speakers in an accentedness rating experiment which is explained in detail in the following section.

2.3 The focus of the present study

It should be clear from the discussion above that most studies of the effect of ethnicity on foreign accentedness perception have looked at Asian speakers, leaving Caucasian NNESs (who are not visually distinguishable from Caucasian NESs constituting the majority in the societies where such studies have been conducted) an under-studied group. However, in the absence of a negative bias, the use of Caucasian speaker participants allows for the testing of other effects of ethnicity, such as an expectation mismatch effect.

In an accentedness rating task in which Caucasian listeners are presented with foreign-accented speech either by itself or together with an Asian or a Caucasian face, reverse linguistic stereotyping may predict a lower foreign accentedness rating for Caucasian NNESs in the audiovisual condition compared to the audio only one in the absence of a negative bias, and a higher foreign accentedness rating for Asian NNESs in the presence of a negative bias (see Table 1). On the other hand, an expectation mismatch effect would predict a similar accentedness score for foreign accented speech presented by itself or with an Asian face and a higher accentedness score for speakers when a Caucasian face is shown.

Table 1: Two accounts' predictions for Asian and Caucasian non-native English speakers' (NNESs) accentedness ratings in two conditions.

	Asian NNESs	**Caucasian NNESs**
Reverse linguistic stereotyping	Audiovisual > Audio	Audiovisual < Audio
Expectation mismatch effect	Audiovisual = Audio	Audiovisual > Audio

This has not yet been fully explored, however, and is the focus of the rest of this paper. In order to distinguish between these two conflicting predictions, I report on an accentedness rating experiment in which 2 groups of non-native speakers of Asian and Caucasian ethnicities were presented to listeners in 3 conditions: audio track of the recording only, video track only, and audiovisual (audio and video tracks of the recording together). Although this study builds on work which combines the same audio clips with different visual stimuli, no such crossing is used in this paper. Instead, the audiovisual condition contains aligned audio and video tracks for each individual speaker recording which is a more naturalistic and ecological design.

3 Method

3.1 Speakers

The speakers in this study were 18 highly proficient but non-native speakers of English (9 L1 Korean and 9 L1 German) and 6 L1 speakers of English (2 New Zealand, 2 Standard American, and 2 Southern British English). All had an age of acquisition of 10 or higher (see Gnevsheva 2015 for more detail). In each language group, half of the participants were males and half females. The age, education, socio-economic class of the participants were comparable to those of the interviewer and listeners: age range = 21–34; average age = 24.875; all were affiliated with the same university in New Zealand at the time of the study (highest academic degree achieved or in progress: 8 Bachelor's, 4 Master's, and 12 PhD).

3.2 Stimuli

The 24 speakers were interviewed by the investigator about their university studies in a quiet room at the university. To elicit spontaneous speech, the speakers were asked to tell the interviewer about the applications of their research or study field. They were recorded with the use of a lapel Opus 55.18 MKII beyerdynamic microphone and an H4n Zoom audio-recorder and a Sony AC-L200C video-recorder, recording at 25 frames per second. The speakers were video-recorded against a plain background with the recorder positioned at their eye-level and the frame including their upper body (see Figure 1). The intensity was scaled to remove variation in volume of the audio-recordings. Short clips of a minimum of 30 words were extracted from the recordings as stimuli. Because

Figure 1: A snapshot from the video track of two non-native English speaking participants.

stopping the clips mid-phrase could have an effect on listeners' perception, complete phrases were used and the exact number of words per clip was allowed to vary (mean length in seconds = 15; range = 8–22). The clips did not contain proper nouns. The audio tracks recorded by the audio-recorder were synchronized with the respective video tracks, so that listeners heard the same track in both the audio only and the audiovisual conditions (see below).

3.3 Listeners

The listeners were 45 Caucasian native speakers of New Zealand English who were recruited through announcements posted around the university campus and via the friend-of-friend method. They were paid a small honorarium for completing the task. 48 people participated in the experiment originally, but 3 participants were excluded from the analysis as they indicated that they had met one or more of the speakers in the experiment. Of the remaining 45, 27 were females and 18 were males with the mean age of 25.47. The listeners were assigned to one of the 3 conditions before meeting with the experimenter: *audio only, audiovisual*, and *video only,* – with 15 participants in each.

3.4 Procedure

The listeners were seated individually in a quiet lab in front of a computer. Stimuli were presented electronically using E-Prime 2.0 (Psychology Software Tools, 2012). The audio stimuli were presented through head-phones; the video stimuli were presented on the computer screen. Before starting the actual task

the listeners read the instructions on the screen, completed a practice trial with a non-linguistic clip which allowed to adjust the volume (in the audio only and audiovisual conditions), and if needed, clarified the procedure with the research assistant. After that, the listeners were presented with 24 clips (1 from each of the 24 speakers) in random order. In the audio only condition, they were presented with the audio clips with a black screen and a fixation point; in the video only condition, they saw the video recordings but did not hear anything, and in the audiovisual condition, they were presented with both the video and the audio signal. In the task, the listeners were instructed to rate the presented clips on a scale which read "No foreign accent at all" and "Very strong foreign accent" at the two extremes using number keys 1 through 7. The listeners could not re-play the clips. At the end, the listeners completed a short biographical questionnaire. The task was self-paced and took up to 30 minutes to complete. The research was reviewed and approved by the University of Canterbury Human Ethics Committee.

4 Results

The accentedness ratings of the NNSs were analyzed in R (R Core Team 2014). A linear mixed-effects model was fit to the NNS data with the perceived accentedness rating as the dependent variable. The maximal model (Barr et al. 2013) included an interaction of *condition* and *L1* as fixed effects; s*peaker*, nested within *L1* group, and *listener* were included as random intercepts; and *L1* as a random slope for *listener* (Table 2). The Korean L1 speakers in the audio condition were chosen as the reference level (Intercept). The *estimate* and the *standard error* columns in the table give us the predicted accentedness rating and standard error for a level respectively. So for the base level (the Korean L1

Table 2: Summary for model of accentedness rating.

	Estimate	Standard error	df	t value	Pr(>|t|)	Significance
(Intercept)	4.415	0.388	30.7	11.371	0.000	–
condition_audiovisual	0.052	0.295	44.3	0.176	0.861	
condition_video	−0.156	0.295	44.3	−0.527	0.601	
L1 German	−0.556	0.504	23.0	−1.101	0.282	
condition_audiovisual:L1G	0.511	0.283	44.1	1.807	0.078	
condition_video:L1G	−0.785	0.283	44.1	−2.775	0.008	**

Note: * $p < 0.05$; ** $p < 0.01$, *** $p < 0.001$.

speakers in the audio condition), the predicted accentedness rating is 4.415. To calculate the predicted accentedness rating for a different level, the respective value in the *estimate* column is added or subtracted. For example, the Korean L1 speakers received a rating 0.052 higher in the audiovisual condition and 0.156 lower in the video condition than in the audio condition. These differences were not significant, as indicated in the *significance* column. This means that the ratings of Korean L1 speakers between the conditions were not significantly different. In the audio condition German and Korean L1 speakers were not rated to be significantly different from each other.

When the model was re-run with levels of L1 re-leveled and German as the Intercept, the accentedness ratings of German L1 speakers' ratings were significantly different between conditions (Table 3). They were rated significantly more accented in the audiovisual condition but less accented in the video condition compared to audio only, which is different from Korean L1 speakers. This interaction of condition and L1 is plotted in Figure 2.

Table 3: Summary for model of accentedness rating (re-leveled).

	Estimate	Standard error	df	t value	Pr(>|t|)	Significance
(Intercept)	3.859	0.377	27.80	10.251	0.000	***
condition_audiovisual	0.563	0.263	44.21	2.143	0.038	*
condition_video	−0.941	0.263	44.21	−3.581	0.001	***
L1K	0.556	0.505	22.99	1.101	0.282	
condition_audiovisual:L1K	−0.511	0.283	44.11	−1.807	0.078	
condition_video:L1K	0.785	0.283	44.11	2.775	0.008	**

Note: * $p < 0.05$; ** $p < 0.01$, *** $p < 0.001$.

To test whether the same NNESs who got a lower score in the video condition also received a higher score in the audiovisual condition compared to audio only, I calculated the means for each speaker in each condition, then for each speaker subtracted the audio mean from the audiovisual mean, obtaining the individual *audiovisual enhancement* score, and the video mean from the audio mean, resulting in the individual *visual accentedness predictability* score. The smaller the audiovisual enhancement score, the more of the visual benefit is found and the less accented the speaker is rated when the visual input is available compared to when it is not. The larger the visual accentedness predictability score, the more "accentless" the speaker looks compared to how he or she sounds. For example, German L1 speaker Lea's mean score across all the

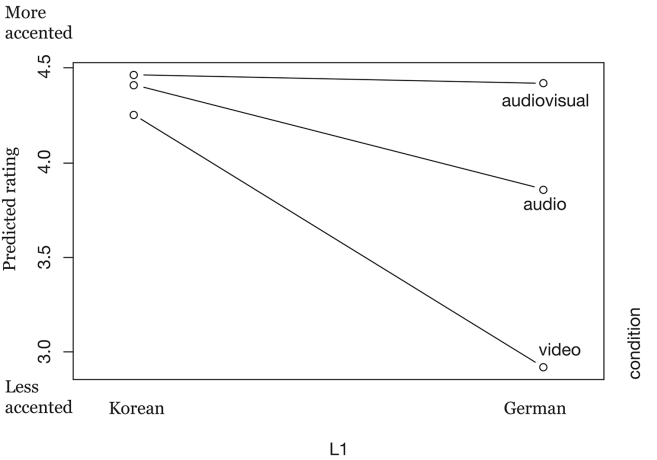

Figure 2: Model prediction for accentedness ratings of Korean and German speakers in the 3 conditions (from model in Table 2).

listeners in the audiovisual condition was 5.80, in the audio condition 5.00, and in the video condition 3.27. Lea's audiovisual enhancement score is 5.80–5.00 = 0.80, and the visual accentedness predictability score is 5.00–3.27 = 1.73. The positive audiovisual enhancement score means that Lea is perceived to be more accented in the audiovisual condition than in the audio only condition. The positive visual accentedness predictability score means that Lea is perceived to be more accented in the audio condition than in the video only one. Calculated in the same fashion, another German L1 speaker Linda's audiovisual enhancement score is 0.93 and visual accentedness predictability score is 1.87. Both of these scores are higher for Linda than for Lea, suggesting that there may be a correlation between them.

To see whether the difference between the audio and the video conditions is predictive of the difference between the audiovisual and audio conditions, I fit a linear regression model with the audiovisual enhancement score as the dependent variable and an interaction between the first language and the visual accentedness predictability score as predictors. However, the interaction was not found to be significant and L1 did not improve model fit, so the final model includes only visual accentedness predictability as an independent variable. In Table 4 we can see that the visual accentedness predictability score was a significant predictor of the audiovisual enhancement score such that the less accented a speaker was rated in the video condition compared to the audio condition the more accented that speaker was rated in the audiovisual condition compared to the audio condition. That is, the less accented a speaker looks, the more accented he/she

Table 4: Summary for model of the audiovisual enhancement score (audiovisual - audio) in accentedness ratings.

	Estimate	Standard error	t value	Pr(>|t|)	Significance
(Intercept)	0.204	0.103	1.983	0.065	–
visual accentedness predictability score (audio – video)	0.189	0.072	2.637	0.018	*

Note: * $p < 0.05$; ** $p < 0.01$, *** $p < 0.001$.

is perceived to be when the video input is available compared to when it is not. This relationship is represented in Figure 3.

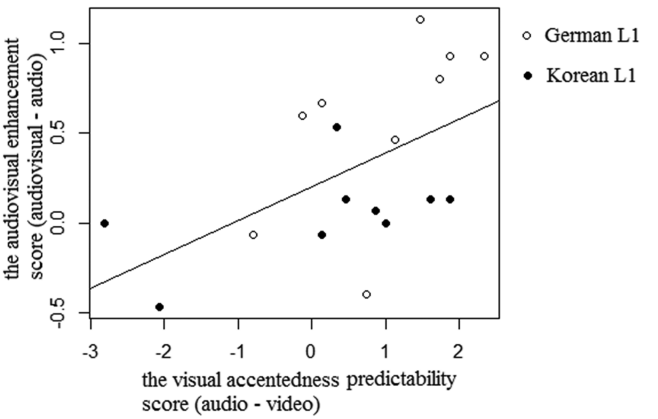

Figure 3: Model prediction for the relationship between the audiovisual enhancement score and the visual accentedness predictability score.

I ran a second linear mixed-effects model to explore the difference in accentedness ratings between German and New Zealand English L1 speakers in the video condition to check whether the Caucasian non-native speakers were rated more accented than the NS based on visual cues only. The model was fit to the German and New Zealand English L1 speaker video condition data with the perceived accentedness rating as the dependent variable. The maximal model included *L1* as a fixed effect, *listener* and *speaker*, nested within *L1* group, as random intercepts, and *L1* as random slope for *listener*. The model illustrates that in the video condition, there was no significant difference between the two language groups (Table 5). This suggests that the listeners were not able to infer the foreign accent based on the video input only.

Table 5: Model summary for accentedness ratings of German and New Zealand English first language (L1) speakers.

	Estimate	Standard error	df	t value	Pr(>|t|)	Significance
(Intercept)	3.233	0.484	11.417	6.686	0.000	–
L1_German	−0.315	0.518	11.039	−0.607	0.556	

5 Discussion

By way of reminder, reverse linguistic stereotyping and expectation mismatch accounts had different predictions for Asian and Caucasian speakers in audio and audiovisual conditions. Reverse linguistic stereotyping predicted a higher accentedness score for Asian speakers and a lower accentedness score for Caucasian speakers in the audiovisual condition. Expectation mismatch effect predicted a similar score in different conditions for Asian speakers and a higher accentedness score in the audiovisual condition for Caucasian speakers.

The accentedness ratings of Korean L1 speakers in the audio condition were not significantly different from the other two conditions, which is different from the findings of Yi et al. (2013). No difference between the audio and the video conditions suggests that the degree of accentedness that the listeners heard in the audio condition was similar to the degree of accentedness they expected to hear from the Asian speakers in the video-only condition. When the video and audio inputs were congruent in the audiovisual condition, as per listeners' expectations, there was no additional effect of ethnicity and the rating in the audiovisual condition was not significantly different from audio only. The negative bias hypothesis, as interpreted by Yi et al. (2013), was not supported as Korean L1 speakers were not rated significantly more accented in the audiovisual condition compared to the audio one. The effect found by Yi et al. (2013) may be due to the experimental design in which listeners were presented with the same sentence and same speakers multiple times.

Based on the results from previous studies, I predicted that the ratings of German L1 speakers in the audiovisual condition would be different from those in the audio condition. The results show that the audiovisual ratings were higher in accentedness scores than the audio only ones. This suggests that reverse linguistic stereotyping did not play the leading role here as a lower accentedness score would be expected. However, as it is likely that the listeners did not expect to hear a foreign accent coupled with a Caucasian face, this could constitute an expectation mismatch effect. That is, listeners were not expecting to hear a foreign accent when they saw a Caucasian speaker, but when they did, the accent was made more

salient, resulting in a higher accentedness score. This interpretation is supported by the significant positive correlation between the difference in the ratings between the audiovisual and audio conditions and between the audio and the video conditions. This means that the more "accentless" the speaker looked and was rated in the video condition compared to the audio condition, the more of a mismatch effect there was and the more the accent "stood out" to the listeners in the audiovisual condition compared to the audio only one.

In accordance with other accounts of reverse linguistic stereotyping (Rubin 1992; Yi et al. 2013), German L1 speakers in the video only condition were rated significantly less accented when the listeners could not hear the speakers as in the audio condition when the accent was actually heard. Moreover, no significant difference was found in the ratings of German and NZE L1 speakers in the video condition, which means that the listeners could not tell the difference between Caucasian L1 and L2 speakers of NZE based on the video input only.

To sum up, Asian NNESs received similar foreign accentedness ratings in the audio and audiovisual conditions while Caucasian NNESs received higher ratings in the audiovisual condition, in line with the predictions of the expectation mismatch effect and contradicting the negative bias hypothesis. These findings support the role of socioindexical expectation described in McGowan (2015) and suggest that reverse linguistic stereotyping may not be the only explanation for an ethnicity effect, but rather a perceived alignment between the audio and the video inputs may have a facilitatory effect while perceived mismatch or misalignment may result in inhibition as the visual and the audio input may be activating conflicting experience-based representations.

This is compatible with the sociophonetic work, described above, which posits experience-based language representations but which has traditionally focused on variation in L1 (e.g., Hay et al. 2006a, who showed that listeners were more likely to make errors in vowel identification when there was a mismatch between actual production and their expected production). When listeners see an Asian speaker, "accented" representations are more likely to be activated, and hearing accented speech reinforces their activation, facilitating easier access and retrieval. However, when listeners see a Caucasian speaker, "accentless" representations are more likely to be activated, but hearing accented speech activates other representations spreading overall activation more thinly and inhibiting access and retrieval.

6 Conclusion

This study explored the effect of socioindexical expectation on accentedness ratings of two groups of NNESs. No differences were found in accentedness

ratings of Korean L1 speakers in the audio, video, and audiovisual conditions, but German L1 speakers were found to be rated significantly less accented in the video condition and more accented in the audiovisual condition compared to the audio only condition. I have argued that this result is due to an expectation mismatch effect for the German L1 speakers. Whereas listeners were expecting to hear an accent when they saw an Asian face paired with accented audio input in the audiovisual condition, they were "surprised" to hear an accent when they saw a Caucasian face, resulting in a higher accentedness rating. Literature on L1 linguistic behavior has often used expectations which are formed by previous experience to explain variation in multiple domains (e.g., Hay et al. 2006b; Niedzielski 1999; discussed above). The current study extends the effect of expectation to NNS perception.

The discussion was framed within usage-based models of speech perception. As listeners' expectations are representative of their past experiences and of societal stereotypes, the current findings may only be applicable to societies with a similar demographic distribution; therefore, it would be interesting to replicate this study in a different setting. In the same setting, quantification of listener experience with Asian NESs and Caucasian NNESs may allow to explore the effect of such experience on accentedness perception ratings in more detail. Future research using Asian NES and NNES listeners may help to clarify whether there is an effect of listener ethnicity on perception of foreign accented speech produced by Asian and Caucasian speakers. Negative bias may be expected to have an effect on perceived accentedness in an expectation mismatch when there is expectation of accentedness based on social characteristics. An expectation mismatch may result in higher accentedness ratings when there is no expectation of accentedness.

These findings suggest that an ethnicity effect may be found for Caucasian speakers as well as Asian speakers, as has been highlighted in previous research. Whereas we may be better aware of stereotyping of minority ethnicities and its effects on people's judgments, we might not be aware of an adverse effect of a majority ethnicity on perceived accentedness to the same extent.

References

Babel, Molly & Jamie Russell. 2015. Expectations and speech intelligibility. *Journal of Acoustical Society of America* 137(5). 2823–2833.

Barr, Dale J., Roger Levy, C. Christoph Scheepers & Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68 (3). 255–278.

Campbell-Kibler, Kathryn. 2007. Accent, (ING), and the social logic of listener perceptions. *American Speech* 82(1). 32–64. doi: 10.1215/00031283-2007-002.

Campbell-Kibler, Kathryn. 2010. Sociolinguistics and perception. *Language and Linguistics Compass* 4(6). 377–389. doi: 10.1111/j.1749-818x.2010.00201.x

Drager, Katie. 2010. Sociophonetic variation in speech perception. *Language and Linguistics Compass* 4(7). 473–480.

Gnevsheva, Ksenia. 2015. Acoustic analysis in Accent of Non-Native English (ANNE) corpus. *International Journal of Learner Corpus Research* 1(2). 256–267.

Han-Gyol, Yi, Rajka Smiljanic & Bharath Chandrasekaran. 2014. The neural processing of foreign-accented speech and its relationship to listener bias. *Frontiers in human neuroscience* 8. 1–12. doi: 10.3389/fnhum.2014.00768

Hay, Jennifer & Katie Drager. 2010. Stuffed toys and speech perception. *Linguistics* 48(4). 865–892.

Hay, Jennifer, Aaron Nolan & K. Katie Drager. 2006a. From fush to feesh: Exemplar priming in speech perception. *The Linguistic Review* 23(3). 351–379. doi: 10.1515/tlr.2006.014.

Hay, Jennifer, Paul Warren & Katie Drager. 2006b. Factors influencing speech perception in the context of a merger-in-progress. *Journal of Phonetics* 34(4). 458–484. doi: 10.1016/j.wocn.2005.10.001.

Johnson, Keith. 1997. Speech perception without speaker normalization: An exemplar model. In Keith Johnson & John W Mullennix eds., *Talker variability in speech processing*, 145–165. San Diego: Academic Press.

Johnson, Keith. 2006. Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *Journal of Phonetics* 34. 485–499. doi: 10.1016/j.wocn.2005.08.004.

Kang, Okim & Donald L Rubin. 2009. Reverse linguistic stereotyping: Measuring the effect of listener expectations on speech evaluation. *Journal of Language and Social Psychology* 28. 441–456.

Kraut, Rachel & Stefanie Wulff. 2013. Foreign-accented speech perception ratings: A multifactorial case study. *Journal of Multilingual and Multicultural Development* 34(3). 249–263. doi: 10.1080/01434632.2013.767340.

Levi, Susannah V., Stephen J Winters & David B Pisoni. 2007. Speaker-independent factors affecting the perception of foreign accent in a second language. *Journal of the Acoustical Society of America* 121(4). 2327–2338.

Lippi-Green, Rosina. 1997. *English with an accent: Language, ideology and discrimination in the United States*. Hoboken: Routledge.

McGowan, Kevin B. 2015. Social expectation improves speech perception in noise. *Language and Speech*. Advance online publication. doi: 10.1177/0023830914565191.

Niedzielski, Nancy. 1999. The effect of social information on the perception of sociolinguistic variables. *Journal of Language and Social Psychology* 18(1). 62–85. doi: 10.1177/0261927x99018001005.

Pierrehumbert, Janet. 2003. Phonetic diversity, statistical learning, and acquisition of phonology. *Language and Speech* 46(2-3). 115–154.

Psychology Software Tools, Inc. [E-Prime 2.0]. 2012. Retrieved from http://www.pstnet.com.

R Core Team. 2014. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. URL http://www.R-project.org/

Rubin, Donald L. 1992. Nonlanguage factors affecting undergraduates' judgments of nonnative English-speaking teaching assistants. *Research in Higher Education* 33(4). 511–531. doi: 10.1007/bf00973770.

Strand, Elizabeth A. 2000. *Gender stereotype effects on speech processing*. Columbus, OH: Ohio State University Dissertation.

Vishnevskaya, Galina. 2008. Foreign accent: Phonetic and communicative hazards. In Ewa Waniek-Klimczak ed., *Issues in accents of English*, 235–251. Newcastle upon Tyne: Cambridge Scholars Publishers.

Yi, Ha-Gyol, Jasmine E.B. Phelps, Rajka Smiljanic & Bharath Chandrasekaran. 2013. Reduced efficiency of audiovisual integration for nonnative speech. *The Journal of the Acoustical Society of America* 134(5). 387–393.